Federated Conformal Predictors for Distributed Uncertainty Quantification

Charles Lu [1 *] Yaodong Yu [2 *] Sai Praneeth Karimireddy [2] Michael I. Jordan [2] Ramesh Raskar [1]

Abstract

Conformal prediction is emerging as a popular paradigm for providing rigorous uncertainty quantification in machine learning since it can be easily applied as a post-processing step to already trained models. In this paper, we extend conformal prediction to the federated learning setting. The main challenge we face is data heterogeneity across the clients — this violates the fundamental tenet of *exchangeability* required for conformal prediction. We propose a weaker notion of *partial exchangeability*, better suited to the FL setting, and use it to develop the Federated Conformal Prediction (FCP) framework. We show FCP enjoys rigorous theoretical guarantees and excellent empirical performance on several computer vision and medical imaging datasets. Our results demonstrate a practical approach to incorporating meaningful uncertainty quantification in distributed and heterogeneous environments. We provide code used in our experiments https://github.com/clu5/federated-conformal.

1. Introduction

Techniques that provide meaningful estimates of uncertainty quantification, such as conformal prediction (Vovk et al., 2005), are critical to deploying machine learning in domains such as healthcare and medical devices (Bhatt et al., 2021; Kompa et al., 2021). *Conformal prediction* adapts a model to output a set of predictions instead of a single maximum likelihood prediction. Furthermore, these prediction sets should contain the correct output with high probability. The uncertainty quantification offered by conformal prediction can increase trust in black-box models for safety-critical decision making (Lu et al., 2022a;c) and has been promoted for several medical applications (Shashikumar et al., 2021; Vazquez & Facelli, 2022; Angelopoulos et al., 2022; Lu et al., 2022b).

In this paper, we consider conformal prediction in the federated learning setting, where several clients (each with some local data distribution $\mathbb{P}_k$) jointly optimize a shared global model on a global distribution ($\sum_{k=1}^{K} \lambda_k \mathbb{P}_k$). Our goal is to provide marginal coverage guarantees on unseen data sampled from the global distribution, $\mathbb{Q}_{\text{test}} = \sum_{k=1}^{K} \lambda_k \mathbb{P}_k$, where λ is a probability vector.

The natural heterogeneity in FL among client distributions $\{\mathbb{P}_k\}$ raises numerous issues. It immediately voids all standard conformal prediction guarantees since it contradicts exchangeability, a fundamental assumption in conformal prediction. It also increases the size of the prediction sets, leading to less useful uncertainty estimates. Additional challenges in FL include potential uncertainty around the global distribution $\mathbb{Q}_{\text{test}}$ (Mohri et al., 2019) and the requirement of a fully distributed implementation with minimal communication (Bonawitz et al., 2022). We show how to carefully design novel methods to overcome these challenges and propose:

1. A framework for federated conformal prediction
2. A theoretical extension of conformal prediction under partially exchangeable client distributions, inexact quantile computations, and uncertain test distributions.
3. A empirical evaluation under data heterogeneity on computer vision and medical imaging datasets.

2. Conformal Prediction

The goal of conformal prediction is to construct a set-valued predictor $\mathcal{C} : \mathcal{X} \to 2^{\mathcal{Y}}$ in such a way as to i) ensure *coverage* at a desired confidence level and ii) improve the *efficiency* by reducing the size of the resulting prediction sets.

Coverage. Suppose we have a multiclass classifier $f : \mathcal{X} \to \Delta^J$ that outputs probability scores for J classes and a desired error rate $\alpha \in (0, 1)$. Then, a prediction set $\mathcal{C}(X) \subseteq 2^{\mathcal{Y}}$ could be constructed by including only those classes where the probability score exceeds the threshold $\tau = 1 - \alpha$ to form the prediction set $\mathcal{C}_\alpha(X) = \{y \in \mathcal{Y} : [f(X)]_y \geq \tau\}$. If we assume these scores are perfectly calibrated, then we expect the true class

[1] MIT Media Lab, Massachusetts Institute of Technology, Cambridge, USA [2] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, USA. Correspondence to: Charles Lu <luchar@mit.edu>, Yaodong Yu <yyu@eecs.berkeley.edu>.

Workshop of Federated Learning and Analytics in Practice, colocated with 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. Copyright 2023 by the author(s).

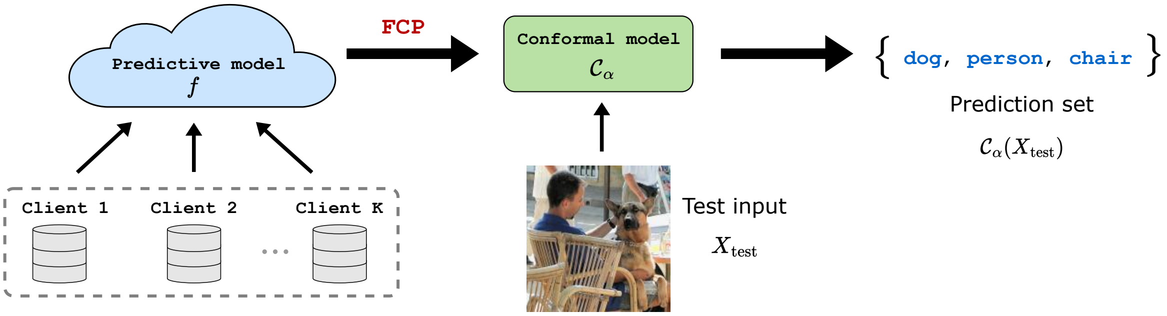


Figure 1. **Overview of federated conformal prediction.** Given K clients and a federated model f, we can obtain a federated conformal model C_α in the distributed environment. The conformal model $\mathcal{C}_\alpha$ produces a prediction set $\mathcal{C}_\alpha(X_{\text{test}})$ for an unseen test sample $(X_{\text{test}}, Y_{\text{test}})$. Prediction sets $\mathcal{C}_\alpha(X_{\text{test}})$ will contain the true label Y_{test} with probability $1-\alpha$, i.e., $\mathbf{P}\left(Y_{\text{test}} \in \mathcal{C}_\alpha(X_{\text{test}})\right) \geq 1-\alpha$. **FCP** represents our proposed federated conformal prediction method. Refer to Algorithm 1 for details on learning federated conformal models.

to be an element of the resulting prediction set with probability at least $1-\alpha$:

$$\mathbf{P}\left(Y \in \mathcal{C}_\alpha\left(X\right)\right) \geq 1-\alpha. \qquad (1)$$

This property is called *marginal coverage*, and predictors enjoy guarantee Eq. (1) are valid conformal predictors.

Efficiency. We prefer a conformal predictor that achieves marginal coverage *efficiently*. A valid predictor is said to be efficient if the expected size of its prediction sets $\mathbb{E}[|C_\alpha\left(X_{n+1}\right)|]$ is small.

Conformal procedure. Deep learning models are known to be poorly calibrated (Guo et al., 2017; Ovadia et al., 2019) that will not automatically be valid conformal predictors with the coverage guarantee. Therefore, to *conformalize* an arbitrary model into a valid conformal predictor, we use the procedure of split conformal prediction to estimate a threshold $\hat{\tau}$ on a *held-out calibration dataset* $\{(X_i, Y_i)\}_{i=1}^n \sim \mathcal{X} \times \mathcal{Y}$ that is assumed to be *exchangeable* with unseen test data $(X_{\text{test}}, Y_{\text{test}})$. A sequence of random variables is exchangeable if the joint probability distribution is identical under any sequence permutation. Exchangeability is a weaker assumption than IID, as exchangeable sequences do not require independence, only identically distributed.

Assume we have some *conformal score function* $S: \mathcal{X} \to \mathbb{R}^+$, where lower values indicate more "conformity" between the test point and the calibration points (see Appendix G for more score functions). One example of a score function is the *least ambiguous set-value classifier* (LAC), defined as $S(X,Y) = 1 - [f(X)]_Y$, where $[f(X)]_Y$ is the softmax score of the true class label. Then, $\hat{\tau}$ can be estimated by taking the $(1-\alpha)$-quantile of conformal scores on the calibration data. The resulting prediction sets $\mathcal{C}_\alpha(X) = \{y \in \mathcal{Y} : S(X,Y) \leq \hat{\tau}\}$ will satisfy Eq. (1) at the desired miscoverage level α.

3. Challenges

Extending conformal prediction to the distributed setting is complicated by the lack of exchangeability between client distributions, reduced efficiency in highly heterogeneous environments, and a communication-efficient distributed implementation of the conformal procedure.

Violation of exchangeability. The main assumption conformal prediction relies upon is exchangeability between the calibration data distribution and the test data distribution during inference. Having random variables be exchangeable implies identical (but not necessarily independent) distributions. However, this assumption of identical distributions is certain to be violated in real-world FL (Terrail et al., 2022). For example, if clients have significant label skew or only have data from some subset of classes, then calibrating on a single client would not provide the correct coverage on other clients (see Figure 6a). Instead, guaranteeing coverage on the global distribution would require tackling non-exchangeable client data distributions.

Decreased efficiency. A conformal predictor is efficient if it achieves marginal coverage with prediction sets that have a small set size. However, data heterogeneity can result in decreased efficiency of conformal predictors (see Figure 6b). For many practical applications, improving the efficiency of federated conformal predictors under data heterogeneity will be a crucial consideration. In our experiments, we show that the choice of FL optimization and score function can have a large influence on FCP efficiency.

Distributed implementation. To ensure correct coverage for non-IID clients, the conformal score quantile needs to be estimated on calibration data from all clients in a distributed manner. However, there are several reasons why a distributed quantile algorithm would be preferable to centrally aggregated quantile estimation. Concerns of privacy underlie the motivation of federated learning (Bonawitz et al., 2022), and sharing the score distribution from all the clients could potentially comprise privacy.

4. Methods

We first introduce the exchangeability assumption in federated learning, which is different from the assumption made in standard conformal prediction methods (Vovk et al., 2005; Lei et al., 2017) and then propose the federated conformal prediction method.

Conformal Prediction with Partial Exchangeability. Recall that we denote the distribution of the k-th client by $\mathbb{P}_k$, i.e., $(X_i^k, Y_i^k) \sim \mathbb{P}_k$, where $\{(X_i^k, Y_i^k)\}_{i=1}^{n_k}$ are the n_k held out calibration samples from the k-th client. Suppose the future test point (X_{n+1}, Y_{n+1}) is sampled from the global distribution $\mathbb{Q}_{\text{test}} = \mathbb{Q}_\lambda$ for some probability vector $\lambda \in \Delta^K$,

$$(X_{n+1}, Y_{n+1}) \overset{\text{i.i.d.}}{\sim} \mathbb{Q}_\lambda, \quad \mathbb{Q}_\lambda = \sum_{k=1}^{K} \lambda_k \mathbb{P}_k. \tag{2}$$

This implies that the global distribution $\mathbb{Q}_\lambda$ is drawn from $\mathbb{P}_k$ with probability λ_k. Hence, with probability λ_k, it has the same distribution as the data points on client k. This forms the basis of our assumption, stated informally below. A more formal version is detailed in Appendix C.1.

Assumption 4.1 (Exchangeability in FL)**.** For a probability vector $\lambda \in \Delta^K$, the softmax scores on client k: $S(X_1^k, Y_1^k), \ldots, S(X_{n_k}^k, Y_{n_k}^k), S(X_{\text{test}}, Y_{\text{test}})$ are exchangeable with probability λ_k.

The above assumption can be interpreted as a variant of *partial exchangeability* (De Finetti, 1980; Diaconis, 1988). This generalization of exchangeability makes no assumptions between $\mathbb{P}_1, \ldots, \mathbb{P}_K$ and does not require independence or identical distributions among the clients.

Theorem 4.2. *Under Assumption 4.1, suppose there are n_k samples from the k-th client, $N = \sum_{k=1}^K n_k$, and $\lambda_k \propto (n_k + 1)$. Define $\hat{q}_\alpha$ as the $\lceil (1-\alpha)(N+K) \rceil$ largest value in $\{(S(X_i^k, Y_i^k))\}_{i\in[n_k],\, k\in[K]}$ and*

$$\mathcal{C}_\alpha(X) = \{y \in \mathcal{Y} : S(X, y) \le \hat{q}_\alpha\}.$$

Then, $C_\alpha(\cdot)$ is a valid conformal predictor with

$$1 - \alpha + \frac{K}{N+K} \ge \mathbf{P}\left(Y_{\textit{test}} \in \mathcal{C}_\alpha(X_{\textit{test}})\right) \ge 1 - \alpha. \tag{3}$$

The proof of Theorem 4.2 can be found in Appendix C.2.

Comparison with the IID setting. Observe that the gap between the upper and lower bound in Eq. (3) is $\frac{K}{N+K} = \frac{1}{N/K+1}$. Thus, the gap depends on the *average* number of data points per client. If a certain client k has very few data points with a small n_k, we will have high uncertainty about the predictions corresponding to client k. This is compensated by client k having a lower weight in our test distribution since $\lambda_k \propto (n_k + 1)$. As a result, our approach is suitable for settings where the average number of data points per client is large. In particular, for non-vacuous bounds, we need

$$\lceil (1-\alpha)(N+K) \rceil \le N \Rightarrow \alpha \ge \frac{1}{N/K+1}.$$

If, instead, we had IID data points, with all data points being exchangeable, Lei et al. (2018) show that we can compute quantiles with a guarantee of

$$(1-\alpha) \le \mathbf{P}\left(Y_{n+1} \in \mathcal{C}_\alpha(X_{n+1})\right) \le \; (1-\alpha) + \frac{1}{N+1}.$$

This results in a $\frac{1}{N+1}$ gap. Our result recovers this as a special case with $K = 1$ but degrades with increasing K. This looseness in the analysis translates to needing a larger quantile. In the IID setting with full exchangeability assumption, we need to pick $\hat{q}_\alpha$ to be the $\lceil (1-\alpha)(N+1) \rceil$ largest value. In contrast, with partial exchangeability, we need the $\lceil (1-\alpha)(N+K) \rceil$ largest value.

Discussion about λ. As described in Theorem 4.2, we assume that $\lambda_k \propto (n_k + 1)$, where λ is a known probability vector for defining the test distribution $\mathbb{Q}_\lambda$. In the federated learning literature, the test distribution is normally defined as the mixture of the distributions of different clients, and the weight of each distribution is proportional to the number of samples from that client (McMahan et al., 2017; Wang et al., 2020). On the other hand, our method can be extended to the setting where λ is unknown during test time (Mohri et al., 2019). Suppose λ is constrained in a convex set Λ, for example, $\Lambda = \{\hat{\lambda} | \hat{\lambda}_k \ge (1-\delta)(n_k+1)/(N+K), k \in [K]\}$, we can modify the definition of $\hat{q}_\alpha$ in our algorithm to achieve valid coverage even when λ violates the assumption $\lambda_k \propto (n_k + 1)$. The modified federated conformal method and the theoretical results can be found in Appendix D.1.

Distributed Quantile Estimation. To learn the set-valued function $\mathcal{C}_\alpha$ in our method, we need to compute the quantile of the conformal scores $\{s_i^k\}_{i\in[n_k],\, k\in[K]}$ that are distributed across K clients, where $s_i^k = S(X_i^k, Y_i^k)$. To reduce the number of communications for estimating the empirical quantiles in our method, we apply distributed quantile estimation techniques (Luo et al., 2016; Dunning, 2021). In particular, we utilize T-Digest, a quantile sketching algorithm well suited for computing online quantile estimates for distributed workflows (Dunning, 2021). We provide an extension of coverage guarantees to approximate quantile estimation methods in Appendix F. We also conduct experiments to study the performance of different quantile estimation methods in Figure 16.

Differential privacy. While the default FL setup does not provide formal privacy guarantees, we can extend our FCP framework to incorporate differential privacy (DP). In particular, we show how to adapt our existing inexact-quantile guarantees, specifically Corollary E.2, to infer coverage under DP. See Subsection F.2 for details.

Table 1. **Design choices for efficient federated conformal predictors.** We conducted several ablation experiments on designing efficient FCP. We found that RAPS, T-Digest, and TCT optimized with cross-entropy to be good defaults for conformal score function, quantile sketcher, and federated optimization procedure, respectively.

FL pretraining method	**FedAvg**	**FedProx**	**TCT**
TCT optimization loss	**Mean squared error**		**cross entropy**
Conformal score function	**LAC**	**APS**	**RAPS**
Distributed quantile estimation	**Mean**	**DDSketch**	**T-Digest**

5. Experiments

Our experiments evaluate the proposed federated conformal prediction (FCP) framework under data heterogeneity with an application of FCP for selective classification. We also perform a number of ablation experiments to evaluate the effect of the conformal score function, quantile estimation method, and FL optimization in Appendix J.

Setup and Datasets. Our experiments used FashionMNIST, CIFAR-10, CIFAR-100, DermaMNIST, PathMNIST, TissueMNIST, and Fitzpatrick17K datasets. For Fitzpatrick17K, We experiment with partitioning disease categories and skin types to model heterogeneous client distributions. In contrast, the clients for the rest of the datasets were partitioned by class (see Appendix H).

For model architectures, we used LeNet (LeCun et al., 1998) for FashionMNIST, Efficient-Net-B1 (Tan & Le, 2019) pretrained on ImageNet (Deng et al., 2009) for Fitzpatrick17k, and ResNet-14 (based off of Idelbayev implementation) for the rest of the datasets.

For each dataset, we trained a centralized baseline and three FL pretraining strategies: FedAvg (McMahan et al., 2016), FedProx (Li et al., 2020), and TCT (Yu et al., 2022) — an approach that first extracts eNTK representations from a FedAvg model before fitting a linear model with SCAFFOLD (Karimireddy et al., 2020). For the decentralized methods, we introduced heterogeneity by partitioning data based on the class between clients (skin type and disease category for the Fitzpatrick17K dataset). We report metrics over 10 random trials, where the calibration and test sets are evenly split. We evaluated three nonconformity scores (defined in Appendix G). See Appendix I for implementation details.

TCT increases efficiency. Our results in Figure 4b and Table 2 show that federated conformal predictors using TCT as the federated pretraining method are more robust to data heterogeneity than other strategies such as FedAvg and FedProx.

Set size correlates with accuracy. In Figure 2, we found a strong negative correlation between prediction set size and prediction accuracy for all datasets. This is intuitive as small sets correspond to more confident predictions while larger sizes correspond to higher conformality scores. This property may be useful for predicting test accuracy on a model without requiring labeled test data.

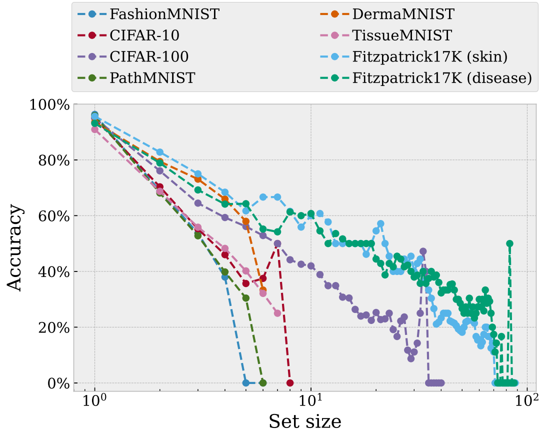


Figure 2. **Set size is negatively correlated with prediction accuracy**. We plotted the median of mean top-1 accuracy for each set size over 100 random trials for each dataset.

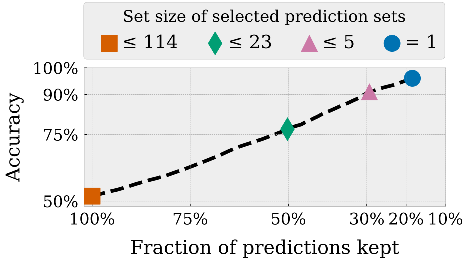
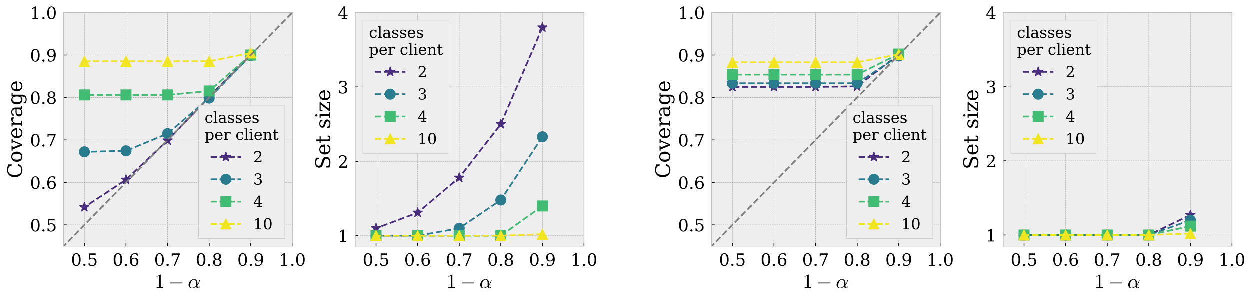


Figure 3. **Selective classification with conformal prediction.** We calibrate a TCT conformal predictor with RAPS score function at $\alpha - 0.1$ on the Fitzpatrick17k skin lesion dataset. We plot the top-1 accuracy as more uncertain predictions are excluded. From a baseline top-1 accuracy of 53%, we can achieve 77% accuracy when filtering out the 50% largest sets.

Selective classification with conformal. In some applications, we may prefer to make fewer, higher-quality predictions instead of predicting every data point. In Figure 3, we plotted the increase in top-1 accuracy that can be achieved by excluding the predictions with a large set size. While this



(a) **Federated Conformal Predictor using FedAvg.**

(b) **Federated Conformal Predictor using TCT.**

Figure 4. **Efficiency of conformal predictors with TCT is more robust under data heterogeneity.** We trained FedAvg and TCT with five clients on CIFAR-10 with four different amounts of data heterogeneity (two, three, four, or all ten classes) for each client. As heterogeneity increases, the average prediction set size also drastically increases for FedAvg but not for TCT.

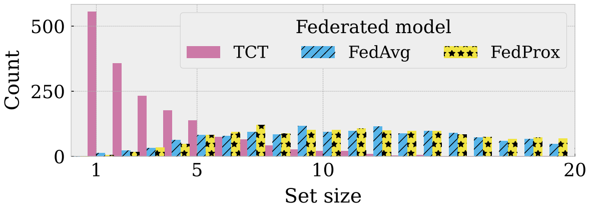


Figure 5. **TCT is more efficient than FedAvg and FedProx even on the same subset of predictions correctly classified by all methods.** We considered only the subset of CIFAR-100 that was correctly predicted by all three federated methods (according to top-1 accuracy) and plotted the distribution of set sizes of their respective prediction sets with LAC at $\alpha = 0.1$ (plot clipped past size 20).

approach is only a heuristic and does not have a coverage guarantee, prediction sets of small size empirically have high coverage (see Table 5), conformal prediction can be extended to provide formal guarantees for controlling types of risks such as false negative rate (Angelopoulos et al., 2021).

Comparing the efficiency of federated conformal predictors under data heterogeneity. In Figure 4b, we compared the coverage and set size of federated conformal predictors using FedAvg and TCT under varying amounts of data heterogeneity. As heterogeneity increases, prediction sets with the FedAvg method have larger set sizes while prediction sets with the TCT maintain small set sizes. In Table 2, we measured the relative increase in average set size over centralized conformal predictors and find that TCT has better efficiency compared to FedAvg and FedProx across all datasets. More detailed results can be found in Table 3.

To better control for differences in predictive performance, we plotted the set size distribution of test examples that were classified correctly by all three decentralized methods in Figure 5. This shows that the better efficiency of TCT is not only a result of greater prediction accuracy but also due to being more calibrated. This advantage in calibration is also robust across the choice of conformity score functions.

Table 2. **Relative inefficiency of decentralized conformal predictors over centralized baseline.** We compare the inefficiency, measured by the ratio of average prediction set size over the centralized baseline, of different decentralized methods for conformal prediction with LAC. Lower inefficiency is better ("1×" would indicate a method is just as efficient as the centralized baseline); bold denotes the most efficient method.

Dataset	$1-\alpha$	FedAvg	FedProx	TCT
FashionMNIST	0.90	8.6×	9.3×	**1.1**×
	0.80	8.0×	7.2×	**1.0**×
CIFAR-10	0.90	3.5×	3.2×	**1.2**×
	0.80	2.5×	2.3×	**1.0**×
CIFAR-100	0.90	3.8×	3.8×	**1.2**×
	0.80	4.1×	4.2×	**1.2**×
DermaMNIST	0.90	3.1×	3.2×	**2.4**×
	0.80	2.1×	2.2×	**1.3**×
PathMNIST	0.90	3.2×	2.8×	**1.2**×
	0.80	2.5×	2.0×	**1.0**×
TissueMNIST	0.90	1.8×	1.8×	**0.9**×
	0.80	2.0×	2.0×	**0.9**×
Fitzpatrick17k (skin type)	0.90	1.3×	1.3×	**1.2**×
	0.80	1.5×	1.4×	**1.2**×
Fitzpatrick17k (disease category)	0.90	2.3×	2.2×	**1.3**×
	0.80	3.1×	2.7×	**1.3**×

6. Conclusion

Conformal prediction is especially well suited to endow federated learning models with finite-sample coverage guarantees that can be used for downstream tasks, e.g., selective classification. This paper introduced conformal prediction to the distributed learning setting with non-IID clients and extended the statistical guarantees of marginal coverage to the mixtures of client distributions in FL. We proposed efficient distributed algorithms to compute these federated conformal predictors (FCP) and extensively evaluated the proposed FCP under data heterogeneity conditions over benchmark computer vision and medical imaging datasets.

References

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *CoRR*, abs/2107.07511, 2021. URL https://arxiv.org/abs/2107.07511.

Angelopoulos, A. N., Bates, S., Malik, J., and Jordan, M. I. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., and Lei, L. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.

Angelopoulos, A. N., Kohli, A. P., Bates, S., Jordan, M., Malik, J., Alshaabi, T., Upadhyayula, S., and Romano, Y. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *International Conference on Machine Learning*, pp. 717–730. PMLR, 2022.

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O., et al. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 401–413, 2021.

Bonawitz, K., Kairouz, P., Mcmahan, B., and Ramage, D. Federated learning and privacy. *Communications of the ACM*, 65(4):90–97, 2022.

De Finetti, B. On the condition of partial exchangeability. *Studies in inductive logic and probability*, 2:193–205, 1980.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Diaconis, P. Recent progress on de Finetti's notions of exchangeability. *Bayesian statistics*, 3:111–125, 1988.

Dunning, T. The t-digest: Efficient estimates of distributions. *Software Impacts*, 7:100049, 2021. ISSN 2665-9638. doi: https://doi.org/10.1016/j.simpa.2020.100049. URL https://www.sciencedirect.com/science/article/pii/S2665963820300403.

Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., and Badri, O. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1820–1828, 2021.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Huang, Z., Wang, L., Yi, K., and Liu, Y. Sampling based algorithms for quantile computation in sensor networks. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pp. 745–756, 2011.

Idelbayev, Y. Proper ResNet implementation for CIFAR10/CIFAR100 in PyTorch. https://github.com/akamaster/pytorch_resnet_cifar10. Accessed: 20xx-xx-xx.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

Kompa, B., Snoek, J., and Beam, A. L. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):4, 2021.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. Distribution-free predictive inference for regression, 2017.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Lu, C., Angelopoulos, A. N., and Pomerantz, S. Improving trustworthiness of ai disease severity rating in medical imaging with ordinal conformal prediction sets. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*, pp. 545–554. Springer, 2022a.

Lu, C., Chang, K., Singh, P., and Kalpathy-Cramer, J. Three applications of conformal prediction for rating breast density in mammography. *arXiv preprint arXiv:2206.12008*, 2022b.

Lu, C., Lemay, A., Chang, K., Höbel, K., and Kalpathy-Cramer, J. Fair conformal predictors for applications in medical imaging. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12008–12016, 2022c.

Luo, G., Wang, L., Yi, K., and Cormode, G. Quantiles over data streams: experimental comparisons, new analyses, and further improvements. *The VLDB Journal*, 25(4): 449–472, 2016.

Masson, C., Rim, J. E., and Lee, H. K. Ddsketch: a fast and fully-mergeable quantile sketch with relative-error guarantees. *Proceedings of the VLDB Endowment*, 12 (12):2195–2205, 2019.

McMahan, H. B., Moore, E., Ramage, D., and y Arcas, B. A. Federated learning of deep networks using model averaging. *CoRR*, abs/1602.05629, 2016. URL http://arxiv.org/abs/1602.05629.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data, 2017.

McSherry, F. and Talwar, K. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pp. 94–103. IEEE, 2007.

Mohri, M., Sivek, G., and Suresh, A. T. Agnostic federated learning. In *International Conference on Machine Learning*, pp. 4615–4625. PMLR, 2019.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Pillutla, K., Laguel, Y., Malick, J., and Harchaoui, Z. Differentially private federated quantiles with the distributed discrete gaussian mechanism. In *International Workshop on Federated Learning: Recent Advances and New Challenges*, 2022.

Romano, Y., Sesia, M., and Candes, E. Classification with valid and adaptive coverage. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3581–3591. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/244edd7e85dc81602b7615cd705545f5-Paper.pdf.

Sadinle, M., Lei, J., and Wasserman, L. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019. doi: 10.1080/01621459.2017.1395341. URL https://doi.org/10.1080/01621459.2017.1395341.

Shashikumar, S. P., Wardi, G., Malhotra, A., and Nemati, S. Artificial intelligence sepsis prediction algorithm learns to say "i don't know". *NPJ digital medicine*, 4(1):134, 2021.

Tan, M. and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Terrail, J. O. d., Ayed, S.-S., Cyffers, E., Grimberg, F., He, C., Loeb, R., Mangold, P., Marchand, T., Marfoq, O., Mushtaq, E., et al. Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. *arXiv preprint arXiv:2210.04620*, 2022.

Vazquez, J. and Facelli, J. C. Conformal prediction in clinical medical sciences. *Journal of Healthcare Informatics Research*, 6(3):241–252, 2022.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Wang, J., Liu, Q., Liang, H., Joshi, G., and Poor, H. V. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., and Ni, B. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Yu, Y., Wei, A., Karimireddy, S. P., Ma, Y., and Jordan, M. I. Tct: Convexifying federated learning using bootstrapped neural tangent kernels, 2022. URL https://arxiv.org/abs/2207.06343.

A. Federated Conformal Prediction Algorithm

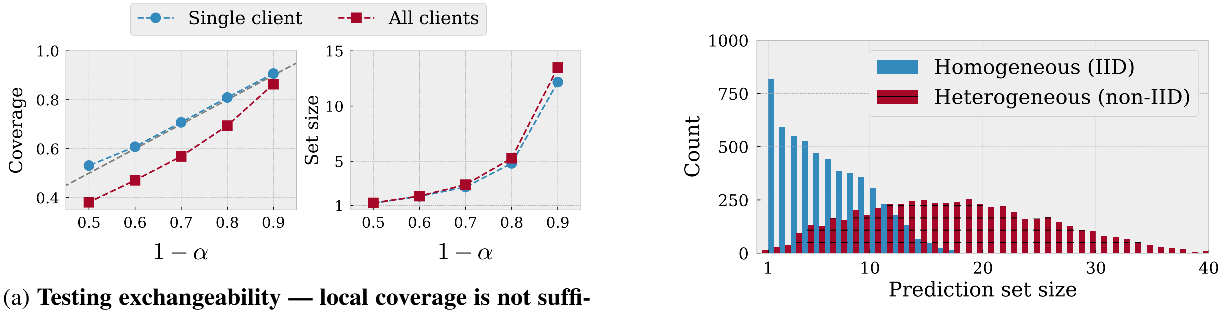
Algorithm 1 Federated Conformal Prediction

Input: global model f_θ parameterized by weights θ, R optimization rounds, K training sets $\{(\hat{X}_i^k, \hat{Y}_i^k)\}_{i=1}^{m_k},\ k \in [K]$, K validation sets $\{(X_i^k, Y_i^k)\}_{i=1}^{n_k},\ k \in [K]$, conformal score function $S : \Delta^J \to \mathbb{R}^+$, error threshold α, federated optimization algorithm FedOpt (we recommend applying TCT (Yu et al., 2022) with logistic regression).

Output: Set-valued function $\mathcal{C}_\alpha(\cdot)$

```
// Step-1:  learn predictive FL model
$f \leftarrow \mathsf{FedOpt}(f_\theta, \{(\hat{X}_i^k, \hat{Y}_i^k)\}_{i\in[m_k],k\in[K]}, R)$
// Step-2:  construct federated conformal predictor
for $k \in \{1, 2, \ldots, K\}$ do
    for $i \in \{1, 2, \ldots, n_k\}$ do
        texttt// Compute conformal score
        $s_i = S(f(X_i), Y_i))$
    end for
    // Sketch and communicate scores
    $\hat{s}_k \leftarrow \mathsf{Sketch}(\{s_i\}_{i=1}^{n_k})$
    Communicate sketch of scores $\hat{s}_k$ to central server
end for
// distributed quantile estimation
$\hat{q}_\alpha := \mathsf{DistributedQuantile}\left(\{\hat{s}_k\}_{k=1}^K, \frac{\lceil(1-\alpha)(N+K)\rceil}{N}\right)$
For $X \in \mathcal{X}, \mathcal{C}_\alpha(X) := \{y \in \mathcal{Y} : S(X, y) \le \hat{q}_\alpha\}$
Return $\mathcal{C}_\alpha(\cdot)$
```

B. Challenges of Federated Conformal Prediction



(a) **Testing exchangeability — local coverage is not sufficient to guarantee global coverage.**

(b) **Non-IID clients degrade efficiency (increases set size).**

Figure 6. We trained ResNet-14 FedAvg models with 20 clients on CIFAR-100. (a) We trained a FedAvg ResNet-14 model on CIFAR-100 with 20 non-IID clients and constructed predictors with the Least Ambiguous set-valued Classifier (LAC) score function on a single client and plotted coverage and set size on the test set for both the first client and all clients. We see marginal coverage is satisfied only for the calibrated client and not across all clients. (b) We trained two FedAvg models, one with IID clients and the other with non-IID clients. We plotted the distribution of predictions by set size and observed that predictors trained with heterogeneous clients have significantly lower efficiency than predictors trained with homogeneous clients.

C. Definitions and Proofs

C.1. Partial exchangeability

We state a more formal version of Assumption 4.1 here.

Assumption 4.1 Define S_i^k to be the ith score on client k i.e. $S_1^k := S(X_1^k, Y_1^k)$, and $\sigma(\cdot)$ to mean the joint probability density function of random variables. Then, we assume there exist a probability vector $\lambda \in \Delta^K$, and random variables

$\{S^k_{n_k+1}\}_{k=1,\ldots,K}$ such that we can decompose the probability density as

$$\sigma(S(X_{\text{test}}, Y_{\text{test}})) = \sum_k \lambda_k \cdot \sigma(S^k_{n_k+1})\,,$$

and further, the joint probability is group-wise permutation invariant i.e. for any set of permutations $\{\pi_1, \ldots, \pi_K\}$,

$$\sigma\left(\begin{bmatrix} S^1_1 & \ldots & S^1_{n_1} & S^1_{n_1+1} \\ \vdots & \ddots & \vdots & \vdots \\ S^K_1 & \ldots & S^K_{n_K} & S^K_{n_K+1} \end{bmatrix}\right) = \sigma\left(\begin{bmatrix} S^1_{\pi_1(1)} & \cdots & S^1_{\pi_1(n_1)} & S^1_{\pi_1(n_1+1)} \\ \vdots & \ddots & \vdots & \vdots \\ S^K_{\pi_K(1)} & \cdots & S^K_{\pi_K(n_K)} & S^K_{\pi_K(n_K+1)} \end{bmatrix}\right).$$

Remark C.1. Consider the example where the client data is sampled independently from $(X^k_i, Y^k_i) \sim \mathbb{P}_{\Bbbk}$ for $i = 1, \ldots, n_k$. Further suppose that the test point is independently drawn from $(X_{\text{test}}, Y_{\text{test}}) \sim \sum_k \lambda_k \mathbb{P}_k$. We will show that this satisfies Assumption 4.1. Note that the scores on client k, $\{S^k_1, \ldots, S^k_{n_k}\}$ are all IID and exchangeable with each other within client k. Now, denote $S^k_{n_k+1} := S(X_{\text{test}}, Y_{\text{test}})|(X_{\text{test}}, Y_{\text{test}}) \sim \mathbb{P}_k$. Then, clearly we have $\{S^k_1, \ldots, S^k_{n_k+1}\}$ are IID and hence exchangeable. Finally, note that $S(X_{\text{test}}, Y_{\text{test}}) = S^k_{n_k+1}$ with probability λ_k.

C.2. Proof of Theorem 4.2

Proof. We denote the total number of samples for conformal calibration as $N = \sum_{k=1}^K n_k$. Given $\lambda_k \propto (n_k+1)$ and $\sum_{k=1}^K \lambda_k = 1$, we have $\sum_{k=1}^K (n_k+1) = N + K$. Therefore,

$$\frac{\lambda_k}{n_k+1} = \frac{1}{N+K}. \tag{4}$$

Meanwhile, for each client k, we define

$$m_k(q) := |\{S(X^k_i, Y^k_i) \le q\}|\,.$$

Recall that we pick $\hat{q}_\alpha$ as the $\lceil (1-\alpha)(N+K) \rceil$-th largest score i.e. it satisfies

$$\sum_{k\in[K]} m_k(\hat{q}_\alpha) = \lceil (1-\alpha)(N+K) \rceil\,. \tag{5}$$

Next, we define the event $\mathcal{E}$ as

$$\mathcal{E} = \Big\{\forall k \in [K], \exists\, \pi_k, (S^k_{\pi_k(1)}, \cdots, S^k_{\pi_k(n_k)}, S^k_{\pi_k(n_k+1)}) = (s^k_1, \cdots, s^k_{n_k}, s^k_{n_k+1})\Big\}\,, \tag{6}$$

where $\{s^k_i\}_{i\in[n_k+1], k\in[K]}$ are the order statistics of the scores i.e., they represent the sorted numerical values of the score values. Note that the index assignment of these values to a particular score is still random and is unconditioned. Upon conditioning of the order statistics, the quantity $m_k(\hat{q}_\alpha)$ is a deterministic quantity.

Then we have

$$\begin{aligned}
&\mathbf{P}\left(S(X_{\text{test}}, Y_{\text{test}}) \le \hat{q}_\alpha \,|\, \mathcal{E}\right) \\
=&\sum_{k=1}^K \lambda_k \cdot \mathbf{P}\left(S(X_{\text{test}}, Y_{\text{test}}) \le \hat{q}_\alpha \;\middle|\; \left\{S(X^k_1, Y^k_1),\, \ldots,\, S(X^k_{n_k}, Y^k_{n_k}),\, S(X_{\text{test}}, Y_{\text{test}})\right\} \text{ are exchangeable}, \mathcal{E}\right) \\
\ge& \sum_{k=1}^K \lambda_k \cdot \frac{m_k(\hat{q}_\alpha)}{n_k+1} \\
=& \frac{\sum_{k=1}^K m_k(\hat{q}_\alpha)}{N+K} \\
=& \frac{\lceil (1-\alpha)(N+K) \rceil}{N+K} \\
\ge& (1-\alpha),
\end{aligned} \tag{7}$$

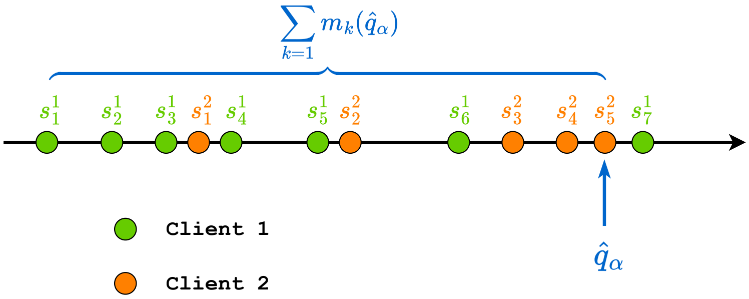


Figure 7. Visualization of the $\hat{q}_\alpha$ of Algorithm 1 with 2 clients and $\alpha = 0.3$, where the 1st client has 7 calibration points and 2nd client has 5 calibration points.

where we apply the partial exchangeable assumption (Assumption 4.1) for the first equality, and the first inequality is by exchangeability given $S(X_1^k, Y_1^k)$, $\ldots$, $S(X_n^k, Y_n^k)$, $S(X_{\text{test}}, Y_{\text{test}})$ are exchangeable random variables. The second equality is because of Eq. (4).

Since Eq. (7) holds for any $(s_1^k, \cdots, s_{n_k}^k, s_{n_k+1}^k)$, $k \in [K]$, we can take expectation on both sides w.r.t. the order statistics and using the towering property of expectations we have

$$\mathbf{P}\left(S(X_{\text{test}}, Y_{\text{test}}) \le \hat{q}_\alpha\right) \ge (1-\alpha). \tag{8}$$

Next, we prove the upper bound of $\mathbf{P}\left(S(X_{\text{test}}, Y_{\text{test}}) \le \hat{q}_\alpha\right)$. Suppose the $\lceil (1-\alpha)(N+K) \rceil$ largest value in $\left\{S\left(X_i^k, Y_i^k\right)\right\}_{i\in[n_k],\, k\in[K]}$ is from the $\hat{k}$-th client, then we have

$$\begin{aligned}
&\mathbf{P}\left(S(X_{\text{test}}, Y_{\text{test}}) \le \hat{q}_\alpha \,|\, \mathcal{E}\right) \\
&\le \sum_{k=1}^{K} \frac{\mathbb{1}\{\hat{k} \ne k\} \cdot \lambda_k \cdot (m_k(\hat{q}_\alpha) + 1) + \mathbb{1}\{\hat{k} = k\} \cdot \lambda_k \cdot m_k(\hat{q}_\alpha)}{n_k + 1} \\
&= \frac{(K-1) + \sum_{k=1}^{K} m_k(\hat{q}_\alpha)}{N+K} \\
&= \frac{(K-1) + \lceil (1-\alpha)(N+K) \rceil}{N+K} \\
&\le (1-\alpha) + \frac{K}{N+K},
\end{aligned} \tag{9}$$

where the second equality is due to Eq (5), and this concludes our proof. □

C.3. Proof of Corollary E.2

Proof. To begin with, we first use $\tilde{q}_\alpha$ to denote ε-approximate $(1-\alpha)$-quantile. Then similar to Eq. (7) in Theorem 4.2, the lower bound of $\mathbf{P}\left(S(X_{n+1}, Y_{n+1}) \le \tilde{q}_\alpha\right)$ is

$$\mathbf{P}\left(S(X_{n+1}, Y_{n+1}) \le \tilde{q}_\alpha | \, \mathcal{E}\,\right) \ge \frac{\sum_{k=1}^{K} m_k(\tilde{q}_\alpha)}{N+K} \ge \frac{(1-\alpha-\varepsilon)(N+K) - 1}{N+K}, \tag{10}$$

where the second inequality is by the Definition E.1 and $\mathcal{E}$ is defined in Eq. (6). Similarly, based on Eq. (9), we can prove the upper bound of $\mathbf{P}\left(S(X_{n+1}, Y_{n+1}) \le \tilde{q}_\alpha\right)$. □

D. Federated Conformal Prediction with Unknown Weights

D.1. Extension to Unknown Weights

Suppose that Assumption 4.1 holds for some arbitrary an *unknown* probability vector $\lambda \in \Lambda \subseteq \Delta^K$. While we may not know the exact weight λ, we know the set Λ it could belong to. Thus, Λ measures our uncertainty about the true weight λ.

Weighted quantile: Pick $\hat{q}_\alpha$ such that

$$\hat{q}_\alpha = \min \left\{ q \,\Big|\, \min_{\hat{\lambda} \in \Lambda} \sum_{k=1}^{K} \frac{\hat{\lambda}_k m_k(q)}{n_k + 1} \geq 1 - \alpha \right\} . \tag{11}$$

From Eq. (7) in the proof of Theorem 4.2, we have almost surely that

$$\begin{aligned} \mathbf{P}\left(S(X_{\text{test}}, Y_{\text{test}}) \leq \hat{q}_\alpha\right) &\geq \sum_{k=1}^{K} \frac{\lambda_k m_k(\hat{q}_\alpha)}{n_k + 1} \\ &\geq \min_{\hat{\lambda} \in \Lambda} \sum_{k=1}^{K} \frac{\hat{\lambda}_k m_k(\hat{q}_\alpha)}{n_k + 1} \\ &\geq (1 - \alpha) . \end{aligned}$$

The second inequality above followed from the assumption that the true $\lambda \in \Lambda$, while the final inequality used the definition of $\hat{q}_\alpha$. We will next consider some important special cases.

Example 1. Consider the case where we have $\lambda_k \approx \frac{n_k+1}{N+K}$. More specifically, suppose there exist $\delta \in [0, 1)$ such that,

$$\lambda_k \geq (1 - \delta) \frac{n_k + 1}{N + K} \quad \forall k \in [K] . \tag{12}$$

This is an approximate version of the setting in Theorem 4.2 where δ controls the approximation factor. Let us construct the set

$$\Lambda = \left\{ \hat{\lambda} \,\big|\, \hat{\lambda}_k \geq (1 - \delta) \tfrac{n_k+1}{N+K} \right\} .$$

Since $\lambda \in \Lambda$, we have

$$\begin{aligned} \mathbf{P}\left(S(X_{\text{test}}, Y_{\text{test}}) \leq q\right) &\geq \min_{\hat{\lambda} \in \Lambda} \sum_{k=1}^{K} \frac{\hat{\lambda}_k m_k(q)}{n_k + 1} \\ &= \min_{\hat{\lambda} \in \Lambda} \sum_{k=1}^{K} \frac{(\hat{\lambda}_k - \frac{n_k+1}{N+K}) m_k(q)}{n_k + 1} + \frac{(\frac{n_k+1}{N+K}) m_k(q)}{n_k + 1} \\ &= \min_{\hat{\lambda} \in \Lambda} \sum_{k=1}^{K} \left(\tfrac{\hat{\lambda}_k (N+K)}{n_k+1} - 1 \right) \frac{m_k(q)}{N + K} + \frac{\sum_k m_k(q)}{N + K} . \end{aligned}$$

Now, by the construction of Λ, we can lower bound the right-hand side above, proving

$$\begin{aligned} \min_{\hat{\lambda} \in \Lambda} \sum_{k=1}^{K} \frac{\hat{\lambda}_k m_k(q)}{n_k + 1} &\geq \sum_{k=1}^{K} (1 - \delta - 1) \frac{m_k(q)}{N + K} + \frac{\sum_k m_k(q)}{N + K} \\ &= (1 - \delta) \frac{\sum_k m_k(q)}{N + K} . \end{aligned} \tag{13}$$

This shows that using the $\lceil (N + K)(1 - \alpha)/(1 - \delta) \rceil$-th largest score as $\hat{q}_\alpha$ suffices to give provable $(1 - \alpha)$ coverage. This neatly recovers the algorithm as well as guarantee when $\lambda_k = \frac{n_k+1}{N+K}$ in Theorem 4.2 by setting $\varepsilon = 0$. Note that using Eq. (11) directly would also yield the required $(1 - \alpha)$ coverage with a smaller value of $\hat{q}_\alpha$ (and hence smaller set sizes). However, this would be less interpretable.

Example 2. Next, consider the special case where we know that $\lambda_k = \frac{1}{K}$ i.e. we want to weight every client equally. In this case, we have

$$\frac{1}{K} \sum_{k=1}^{K} \frac{m_k(\hat{q}_\alpha)}{n_k + 1} \leq \mathbf{P}\left(S(X_{\text{test}}, Y_{\text{test}}) \leq \hat{q}_\alpha\right) \leq \frac{1}{K} \sum_{k=1}^{K} \frac{m_k(\hat{q}_\alpha) + 1}{n_k + 1} .$$

The gap between the upper and lower bounds, in this case, is $\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k+1}$. This is the inverse of the *harmonic mean* of $\{(n_1+1),\dots,(n_K+1)\}$. Thus,

$$\max_k \frac{1}{n_k+1} \geq \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k+1} \geq \frac{K}{\sum_k (n_k+1)} = \frac{1}{N/K+1}\,,$$

with equalities holding only if all clients have equal data with $n_k = \frac{K}{N}$. Thus, the gap when $\lambda_k = \frac{1}{K}$ is larger than when we have $\lambda_k \propto (n_k+1)$ and is more sensitive to clients with little data. However, it is still better than using a client's data on its own, which would have a gap of $\max_k \frac{1}{n_k+1}$.

D.2. Unified Analysis of Federated Conformal Method

In this subsection, we present a unified analysis of our proposed conformal prediction by considering the following three factors: (a) K number of clients; (b) δ error parameter of λ; (c) ε error parameter of distributed quantile estimation.

Theorem D.1. *Suppose there are n_k samples from the k-th client and $\lambda_k \propto (n_k+1)$, if we define $\hat{q}_\alpha$ as the $\lceil (1-\alpha)(N+1)\rceil$ largest value in $\{(S(X_i^k, Y_i^k))\}_{i\in[n_k],\,k\in[K]}$. Then for $(X_{test}, Y_{test}) \sim \mathbb{Q}_\lambda$, we have*

$$\mathbb{E}\left[\mathbb{1}\{Y_{test} \in \mathcal{C}_\alpha(X_{test})\}\right] \geq (1-\delta)\frac{(1-\alpha-\varepsilon)(N+1) - \mathbb{1}\{\varepsilon > 0\}}{N+K}, \tag{14}$$

where $\delta \in [0,1)$ is the error parameter of λ defined in Eq. (12), $\varepsilon \in [0,1)$ is the error parameter of distributed quantile estimation defined in Definition E.1.

Proof. To start with, by Eq. (13), we have

$$\min_{\hat{\lambda}\in\Lambda}\sum_{k=1}^{K}\frac{\hat{\lambda}_k m_k(\hat{q}_\alpha)}{n_k+1} \geq (1-\delta)\frac{\sum_k m_k(\hat{q}_\alpha)}{N+K}\,, \tag{15}$$

then by applying the Definition E.1, we have

$$\sum_k m_k(\hat{q}_\alpha) \geq (1-\alpha-\varepsilon)(N+1) - \mathbb{1}\{\varepsilon > 0\}, \tag{16}$$

therefore, we have

$$\mathbb{P}\left(S(X_{\text{test}}, Y_{\text{test}}) \leq \hat{q}_\alpha\right) \geq \min_{\hat{\lambda}\in\Lambda}\sum_{k=1}^{K}\frac{\hat{\lambda}_k m_k(\hat{q}_\alpha)}{n_k+1} \geq (1-\delta)\frac{(1-\alpha-\varepsilon)(N+1) - \mathbb{1}\{\varepsilon > 0\}}{N+K}, \tag{17}$$

which concludes our proof. □

Remark D.2. When $\{(S(X_i^k, Y_i^k))\}_{i\in[n_k],\,k\in[K]}$ and $S(X_{\text{test}}, Y_{\text{test}})$ are exchangeable, then it is equivalent to the setting where $K = 1$. In this scenario, when $\delta = \varepsilon = 0$, Eq. (14) is the same as the standard coverage guarantee of non-FL conformal prediction methods.

E. Distributed Quantile Estimation

In order to estimate the empirical quantile of the conformal score distribution using the clients' calibration dataset, we utilize T-Digest, a quantile sketching algorithm suited for distributed workflows (Dunning, 2021). T-Digest is a probabilistic data structure based on 1D K-means that computes online quantile estimates. Importantly, this data structure is mergeable, which allows for distributed aggregation in parallel workflows.

As the quantile is approximately computed in the distributed quantile estimation methods, in what follows, we provide the coverage guarantees of our method when using inexact quantiles. We first introduce the ε-approximate β-quantile (Luo et al., 2016).

Definition E.1 (ε-approximate β-quantile)**.** For an error $\varepsilon \in (0,1)$, the ε-approximate β-quantile is any element with rank between $(\beta-\varepsilon)N$ and $(\beta+\varepsilon)N$, where N is the total number of elements.

If the distributed quantile estimation method outputs ε-approximate β-quantiles, then our method achieves desired coverage approximately.

Corollary E.2. *Under Assumption 4.1, suppose there are n_k samples from the k-th client, $N = \sum_{k=1}^{K} n_k$, and $\lambda_k \propto (n_k+1)$, and the sketch algorithm outputs the ε-approximate $(1-\alpha)$-quantile. Then the output $\mathcal{C}_\alpha$ of Algorithm 1 satisfies*

$$\begin{aligned} \mathbf{P}\left(Y \in \mathcal{C}_\alpha(X)\right) &\geq 1 - \alpha - \hat{\varepsilon} - \frac{\mathbb{1}\{\varepsilon > 0\}}{N+K}, \\ \mathbf{P}\left(Y \in \mathcal{C}_\alpha(X)\right) &\leq 1 - \alpha + \hat{\varepsilon} + \frac{K}{N+K}, \end{aligned} \tag{18}$$

where $\hat{\varepsilon} = \varepsilon N/(N+K)$.

As suggested by the above result, when ε is small, Algorithm 1 achieves similar coverage compared to Theorem 4.2 with the exact quantile. When the approximation error $\varepsilon = 0$, the results in Corollary E.2 reduces to the ones in Theorem 4.2. To elucidate the above theoretical results further, we take the algorithm from (Huang et al., 2011) as an example and provide a detailed theoretical statement in Appendix F.

We also conduct experiments to study the performance of different quantile estimation methods in our algorithm. The results are summarized in Figure 16, and we find that T-digest achieves similar performance as the exact quantile. Specifically, for T-digest, the accuracy of estimating the q-quantile is approximately constant relative to $q \cdot (1-q)$, and memory of $\Theta(\delta)$, where δ is a compression parameter $\delta \ll n$. This relative error bound is well-suited for computing quantiles near 0 or 1, which is often necessary with small α values, compared to absolute error bounds of other quantile approximation methods such as DDSketch (Masson et al., 2019). In other words, T-digest achieves smaller $\hat{\varepsilon}$ in Eq. (18) when α is close to 1.

As shown in Theorem 3 of Huang et al. (2011), computing an ε-approximate $(1-\alpha)$-quantile requires at most $O(K/\varepsilon)$ bits to be communicated to the server. Therefore, we can restate Corollary E.2 incorporating this guarantee.

Example 1. Suppose we have K clients, and there are n_k samples from the k-th client and $\lambda_k = (n_k+1)/(N+K)$, where $N = \sum_{k=1}^{K} n_k$ is the total number of calibration samples. Suppose $N \geq K$ and $\alpha < 1/(N+K)$. Then, let us set $\varepsilon = ((N+K)\alpha - 1)/N$. For this value of ε, the distributed quantile estimation algorithm takes $O(K/\varepsilon)$ bits of communication and guarantees that the output C_α of Algorithm 1 satisfies $\mathbf{P}\left(Y \in \mathcal{C}_\alpha(X)\right) \geq 1 - 2\alpha$.

F. Additional Results of Distributed Quantile Estimate

F.1. Communication required

In this section, we take the algorithm from Huang et al. (2011) as an example and provide a detailed theoretical statement below.

As shown in Theorem 3 of Huang et al. (2011), computing an ε-approximate $(1-\alpha)$-quantile requires at most $O(K/\varepsilon)$ bits to be communicated to the server. Therefore, incorporating this guarantee, we can restate Corollary E.2.

Example 1. Suppose we have K clients, and there are n_k samples from the k-th client and $\lambda_k = (n_k+1)/(N+K)$, where $N = \sum_{k=1}^{K} n_k$ is the total number of calibration samples. Suppose $N \geq K$ and $\alpha < 1/(N+K)$. Then, let us set $\varepsilon = ((N+K)\alpha - 1)/N$. For this value of ε, the distributed quantile estimation algorithm takes $O(K/\varepsilon)$ bits of communication and guarantees that the output C_α of Algorithm 1 satisfies $\mathbf{P}\left(Y \in \mathcal{C}_\alpha(X)\right) \geq 1 - 2\alpha$.

F.2. Differential privacy

We leverage the exponential mechanism for selecting a private quantile (McSherry & Talwar, 2007). Assume we are given a set of scores $(s_1, \ldots, s_N)$ and a privacy parameter δ is chosen (while typically ε is used to indicate the privacy parameter, here we use δ to avoid notation clash with ϵ from Corollary E.2). Define the following utility function:

$$u(x) = -\left||\{i: \text{ s.t. } s_i \leq x\}| - N(1-\alpha)\right|, \quad \text{defined for } x \in (s_1, \ldots, s_N).$$

Its sensitivity can clearly be seen to be 1. Thus, to achieve δ-DP, we need to output a quantile x such that

$$\Pr(x = s_i) \propto \exp(\delta u(s_i)/2).$$

This is equivalent to the following procedure:

1. Pick $r \in [N]$ such that $\Pr(r = i) \propto \exp(-\delta|i - (1-\alpha)N|/2)$
2. Return the r-th largest number in $\{s_1, \ldots, s_N\}$ using an exact distributed quantile estimator.

Thus, by the exponential mechanism (McSherry & Talwar, 2007), the output of the above mechanism satisfies δ-DP. Further, by examining the tail of the exponential distribution, we have that $\beta \in [1 - \alpha \pm O(\frac{-\log(\delta\alpha)}{\delta N})]$ with probability at least $1-\alpha$. This satisfies the ϵ approximate quantile estimator for $\epsilon = \frac{1}{\delta N}$ as in Definition E.1. Combining these results with Corollary E.2, we get the following result.

Corollary F.1. *The exponential mechanism returns a quantile-estimator that satisfies δ-DP and has a coverage guarantee of*

$$\mathbf{P}\left(Y \in \mathcal{C}_\alpha(X)\right) \geq 1 - 2\alpha - O\left(\frac{\log\left(\frac{1}{\delta\alpha}\right)}{\delta N}\right) - \frac{\mathbb{1}\{\varepsilon > 0\}}{N+K}\,.$$

A similar coverage guarantee can be obtained with stronger *local* DP guarantees by relying on the privately distributed quantile estimator such as those in Pillutla et al. (2022).

G. Nonconformity score functions

Given a trained classifier f, a score function S, and an exchangeable calibration set of features $X \in \mathbb{R}^d$ with labels $Y \in \mathcal{Y} = \{1, 2, \ldots, J\}$, a prediction set that outputs a subset of classes $2^{\mathcal{Y}}$ can be formed by

$$\mathcal{C}(X) = \{y \in \mathcal{Y} : S(X, y) \leq \hat{q}_\alpha\}\,, \tag{19}$$

where $\hat{q}_\alpha$ is the $(1-\alpha)$-empirical[1] quantile of the calibration scores $\hat{q}_\alpha = \text{Quantile}(\{s_1, s_2, \ldots, s_n\}, \frac{\lceil (n+1)(1-\alpha) \rceil}{n})$. We consider three commonly used score functions for conformal prediction classification tasks in our experiments:

1. *least ambiguous set-valued classifiers* (LAC) (Sadinle et al., 2019)

   $$S_{\text{LAC}}(x, y) = 1 - [f(x)]_y\,, \tag{20}$$

   where $[f(x)]_y$ indexes the score of the true label,

2. *adaptive prediction sets* (APS) (Romano et al., 2020)

   $$S_{\text{APS}}(x, y) = \sum_{c=1}^{C} \left[\pi\left(f(x)\right)\right]_c\,, \tag{21}$$

   where $C = \sup\left\{j \in J : \sum_{i=1}^{j} [\pi(f(x))]_i \leq 1 - \alpha\right\}$, and $\pi\left(f(x)\right)$ is the permutation that sorts the scores in descending order,

3. and *regularized adaptive prediction sets* (RAPS) (Angelopoulos et al., 2020)

   $$S_{\text{RAPS}}(x, y) = \sum_{c=1}^{C} \left[\pi\left(f(x)\right) + a \cdot \mathbb{1}[\{c > b\}]\right]_c\,, \tag{22}$$

   where $\mathbb{1}\{\cdot\}$ is the indicator function, C is defined same as above, and (a, b) are regularization parameters.

[1] Essentially the $1-\alpha$ quantile with a small correction.

H. Dataset Information

The Fitzpatrick17K skin lesion dataset is a challenging real-world dataset with 114 different skin conditions with reported benchmark top-1 accuracy of 20.3% in the original paper (Groh et al., 2021). The 114 skin conditions are grouped into three broad disease categories: "non-neoplastic", "malignant", and "benign". Additionally, each image is labeled with a Fitzpatrick skin type, which is rated on an ordinal six-point scale that approximately measures skin color (lighter skin tones have higher rates of skin cancer; see Figure 9).

For FashionMNIST (Xiao et al., 2017), one class was assigned to each of the 10 clients. For CIFAR10 (**?**), two classes were assigned to each of the five clients. For CIFAR100, five classes were assigned to each of the 20 clients. For MedMNIST (Yang et al., 2023) datasets (DermaMNIST, PathMNIST, TissueMNIST) were partitioned by groups of 2 and 3 classes; see Appendix H for the exact partition.

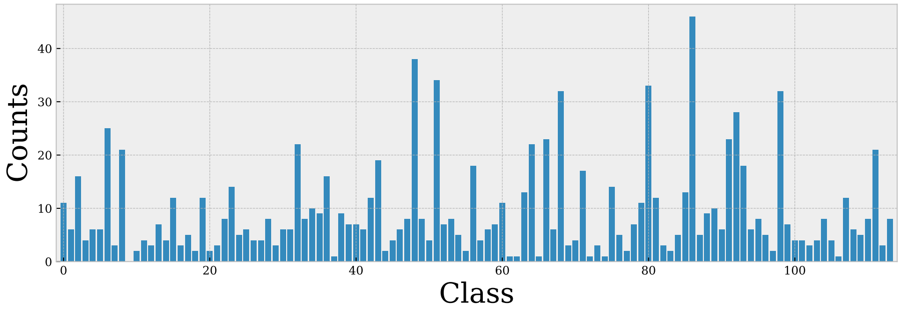


Figure 8. **Class distribution of Fitzpatrick17k skin lesion dataset.** The Fitzpatrick17K dataset contains 16,577 photography images collected from two dermatology atlases with labels for 114 different skin conditions.

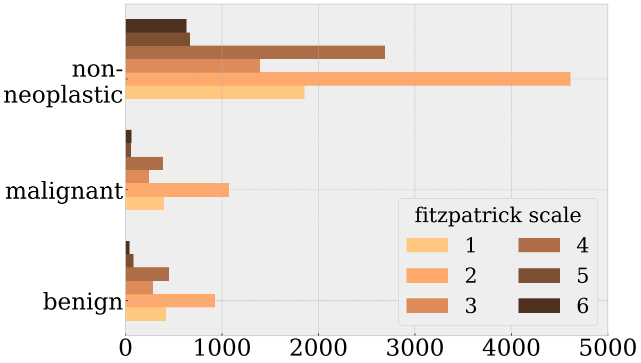


Figure 9. **Distribution of skin types of Fitzpatrick17k skin lesion dataset**. Most images in the Fitzpatrick17k are also labeled with their Fitzpatrick skin type, which measures the amount of melanin pigment in the skin.

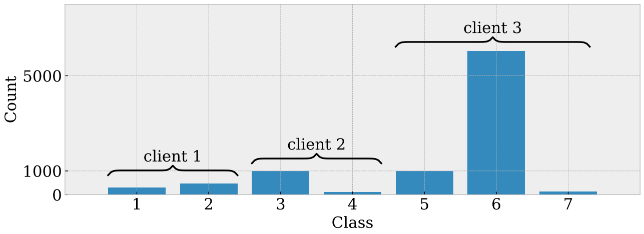


Figure 10. **Class distribution and client partition on DermaMNIST dataset.** DermaMNIST is a dermatoscopy dataset with 10,015 images labeled with one of 7 conditions.

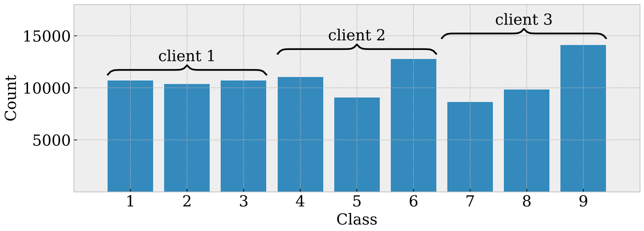


Figure 11. **Class distribution and client partition on PathMNIST dataset.** PathMNIST is a colon pathology dataset with 107,180 images labeled with one of 9 conditions.

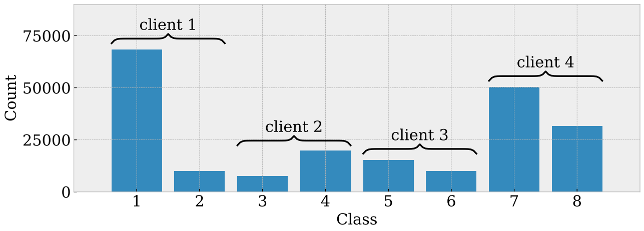


Figure 12. **Class distribution and client partition on TissueMNIST dataset.** TissueMNIST is a kidney cortex microscope dataset of 236,386 images labeled with one of 8 conditions.

I. Implementation Details

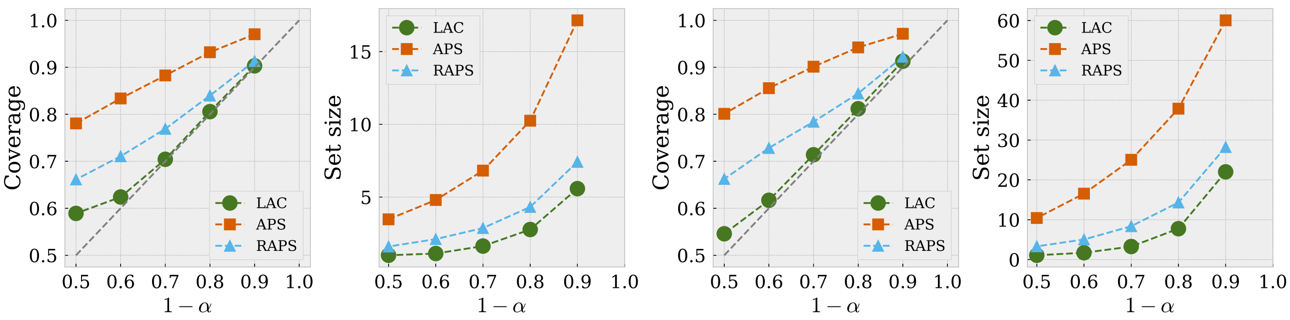
For model architectures, we used LeNet (LeCun et al., 1998) for FashionMNIST, Efficient-Net-B1 (Tan & Le, 2019) pretrained on ImageNet (Deng et al., 2009) for Fitzpatrick17k, and ResNet-14 (based off of Idelbayev implementation) for the rest of the datasets.

We use the TCT approach proposed by Yu et al. (2022) to train a distributed model suited for data heterogeneity. TCT has two optimization stages: a pretraining stage that trains a FedAvg model and a second stage that solves a convex approximation of the trained model. We found the squared loss used in the paper resulted in poor calibration and instead used a cross-entropy loss in the second stage. For FedAvg, 200 communication rounds with five local epochs. For stage 2 of TCT, we take the FedAvg model trained with 100 epochs (instead of 200) and additionally train 100 communication rounds with a learning rate of 0.0001 and 500 local steps.

For the Fitzpatrick17k dataset, we use the Torchvision implementation of the EfficientNet architecture pre-trained on Imagenet-V1. For all other datasets, we use ResNet-14. We train all models with SGD with 0.9 momentum, a learning rate of 0.01 (0.001 for Fitzpatrick17k), and a minibatch size of 64. Data augmentation, such as random flipping and cropping, was applied to all datasets during training; for the Fitzpatrick17k dataset, random color jittering and rotations were also applied.

For all conformal score functions, we forced each prediction set to contain at least one prediction by always including the class with the highest score in the prediction set. As a consequence, coverage will never fall below top-1 classification accuracy and empirical coverage may exceed the marginal coverage guarantee. If empty prediction sets are allowed, the upper bound of marginal coverage holds. For RAPS, we set the regularization parameters $a = 1$ and $b = 0.001$ ($b = 0.00001$ for the Fitzpatrick17k dataset).

J. Additional Experimental Results



(a) **CIFAR-100**

(b) **Fitzpatrick17K**

Figure 13. **LAC is the most efficient conformal score function.** When disallowing empty prediction sets, adaptive score functions such as APS and RAPS have more coverage than necessary than the marginal guarantee. LAC has tighter coverage and a smaller average set size than APS and RAPS.

Comparing score functions. We can approximately evaluate conditional coverage as size-stratified coverage (SSC) as proposed by Angelopoulos et al. (2020): In some sense, this measures a predictor's adaptiveness to different inputs, meaning that larger sets represent more difficult or uncertain predictions and smaller sets represent easier or more confident predictions (Angelopoulos & Bates, 2021). We fixed TCT as our model and evaluate different choices of conformal score function. We observed that RAPS has tighter coverage when stratified by set size, shown in Table 5, while LAC has too much coverage on small sets and too little coverage on large sets.

Comparing quantile methods. In Figure 16, we compared four different methods of computing approximate quantiles on Fitzpatrick17k with TCT and LAC. Simply averaging client quantiles produces a biased estimate which does not provide correct coverage while DDSketch, another distributed quantile estimator, results in bad quantile estimates at high confidence thresholds. Only T-Digest produces the expected coverage of the centralized quantile across α thresholds.

Impact of FL optimization. In Table 6, we compared squared loss and cross-entropy loss for TCT optimization and found that while the squared loss model had higher top-1 classification accuracy, the resulting prediction sets were much larger compared to models optimized with cross-entropy. We further investigated the difference in calibration error between losses

Table 3. **Results of conformal prediction with different models.** We report coverage and size of prediction sets calibrated centrally and with our decentralized framework (FedAvg, FedProx, and TCT). The mean over 10 random trials is reported with standard deviations in the range of ± 0.01 for coverage results and ± 0.1 for size. Smaller set sizes indicate more efficient prediction sets.

			FedAvg		FedProx		TCT		Centralized	
Dataset	classes	$1-\alpha$	coverage	size	coverage	size	coverage	size	coverage	size
FashionMNIST	10	0.90	0.90	3.2	0.90	2.2	0.91	1.2	0.90	1.1
		0.80	0.80	2.2	0.80	1.5	0.89	1.1	0.88	1.0
		0.70	0.70	1.3	0.70	1.2	0.89	1.0	0.88	1.0
CIFAR-10	10	0.90	0.90	3.8	0.90	3.5	0.90	1.3	0.90	1.1
		0.80	0.80	2.5	0.80	2.3	0.82	1.0	0.88	1.0
		0.70	0.70	1.8	0.70	1.6	0.82	1.0	0.88	1.0
CIFAR-100	100	0.90	0.90	17.3	0.90	17.9	0.90	5.6	0.90	4.8
		0.80	0.80	9.2	0.80	9.9	0.80	2.8	0.80	2.4
		0.70	0.70	5.4	0.70	6.1	0.70	1.6	0.70	1.4
DermaMNIST	7	0.90	0.90	3.1	0.90	3.2	0.90	2.4	0.90	1.7
		0.80	0.81	2.1	0.80	2.2	0.81	1.3	0.81	1.2
		0.70	0.73	1.5	0.71	1.5	0.74	1.0	0.75	1.0
PathMNIST	9	0.90	0.90	3.2	0.90	2.8	0.90	1.2	0.90	1.0
		0.80	0.80	2.5	0.80	2.0	0.84	1.0	0.89	1.0
		0.70	0.70	2.0	0.70	1.5	0.84	1.0	0.89	1.0
TissueMNIST	8	0.90	0.90	5.3	0.90	5.2	0.90	2.7	0.90	3.0
		0.80	0.80	4.2	0.80	4.1	0.80	1.8	0.80	2.0
		0.70	0.70	3.3	0.70	3.3	0.70	1.3	0.70	1.5
Fitzpatrick17K (disease partition)	114	0.90	0.91	24.9	0.91	25.6	0.91	20.0	0.90	16.1
		0.80	0.79	8.6	0.80	8.7	0.80	6.0	0.81	6.1
		0.70	0.69	3.8	0.69	3.8	0.70	2.8	0.71	2.7
Fitzpatrick17K (skin type partition)	114	0.90	0.91	22.8	0.91	22.6	0.91	21.3	0.90	18.2
		0.80	0.81	9.2	0.81	9.7	0.80	7.4	0.80	6.3
		0.70	0.69	4.1	0.72	4.5	0.70	3.2	0.70	2.6

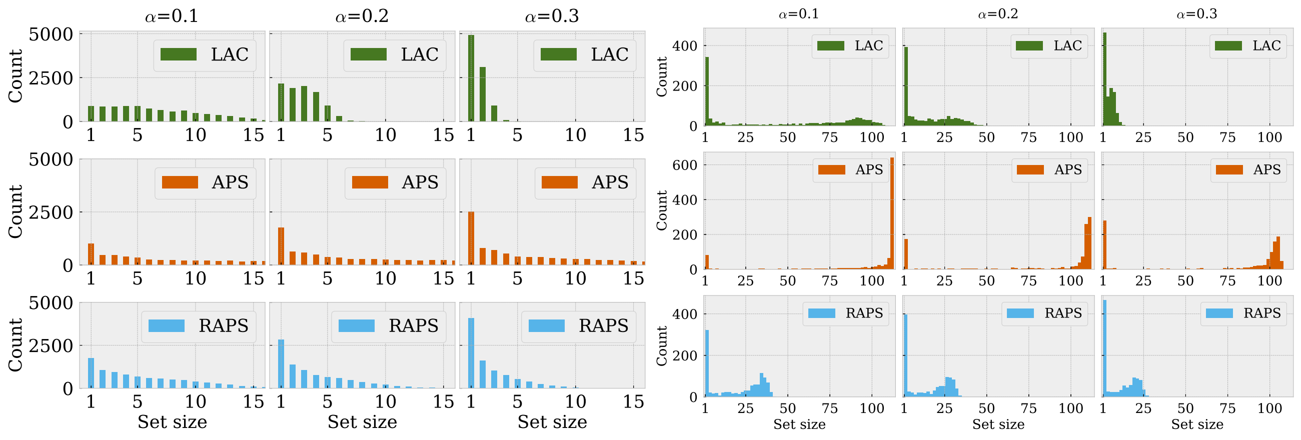
in Figure 17 and found that models optimized with squared loss result in extremely small softmax values, which have poor calibration even after temperature scaling.

Table 4. **Efficiency gain of TCT is robust across the choice of the score function.** Comparing the mean size of prediction sets at $\alpha = 0.1$ of CIFAR-100 test examples that were correctly predicted by top-1 across all three models (TCT, FedAvg, FedProx). All methods achieve perfect coverage on this subset.

	LAC	APS	RAPS
METHOD	SIZE	SIZE	SIZE
FEDAVG	13.2	13.3	7.7
FEDPROX	12.9	14.2	8.1
TCT	3.2	7.6	2.4

Table 5. **RAPS is more adaptive than LAC and APS.** An adaptive conformal predictor outputs larger sets when the predictor is highly uncertain and smaller sets when the predictor is highly confident. We evaluated the size conditional coverage and average set size of each quartile of set sizes and see that LAC has lower than $1 - \alpha$ coverage on prediction with larger sizes while RAPS has tighter $1 - \alpha$ coverage across quartiles.

	SIZE PERCENTILE	0 – 25		25 – 50		50 – 75		75 – 100	
DATASET	SCORE FUNCTION	COVERAGE	SIZE	COVERAGE	SIZE	COVERAGE	SIZE	COVERAGE	SIZE
CIFAR-100	LAC	0.97	2.0	0.93	4.5	0.86	6.9	0.81	10.1
	APS	0.96	2.1	0.97	8.4	0.97	19.3	0.96	37.6
	RAPS	0.94	1.3	0.92	3.8	0.89	8.3	0.88	18.6
FITZPATRICK (DISEASE)	LAC	0.96	2.5	0.89	9.7	0.86	20.7	0.84	37.8
	APS	0.94	7.0	0.97	45.6	0.97	74.5	0.97	89.5
	RAPS	0.92	1.2	0.87	10.2	0.90	32.8	0.92	56.9



(a) **CIFAR-100**

(b) **Fitzpatrick17K**

Figure 14. **Comparing the distribution of prediction size between conformal score functions.** We trained an FL model with TCT on CIFAR-100 with 20 heterogeneous clients and plotted the number of predictions at each set size. With the Fitzpatrick17k dataset, APS is extremely inefficient at achieving coverage; most prediction sets are close to 114, the maximum number of possible classes.

Table 6. **Optimizing TCT with squared loss results in inefficient conformal predictors.** Comparing the average set size of TCT optimized with different stage 2 loss functions on CIFAR-100 at $\alpha = 0.1$ across conformal score functions (LAC, APS, and RAPS).

LOSS FUNCTION	ACCURACY	LAC	APS	RAPS
CROSS-ENTROPY	61%	6.5	18.0	8.3
SQUARED LOSS	58%	12.1	78.8	22.5

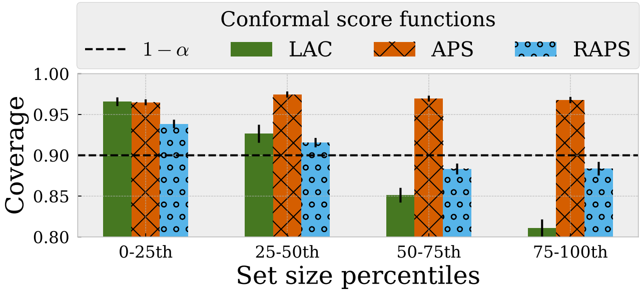


Figure 15. **RAPS score function maintains tighter coverage at different set sizes than LAC and APS.** We evaluated the size conditional coverage and average set size of each quartile of set sizes and see that LAC has lower than $1-\alpha$ coverage on prediction with larger sizes while RAPS has tighter $1-\alpha$ coverage across different set sizes.

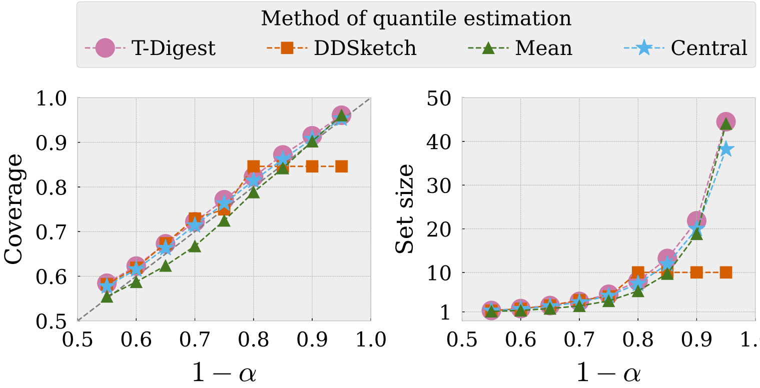


Figure 16. **Comparing different methods of distributed quantile approximation on Fitzpatrick17k.** Naively averaging client quantiles produces a biased estimate that does not provide correct coverage at lower thresholds (Mean). DDSketch has large errors computing quantiles at high α values. Only T-Digest closely approximates the true quantile of the centralized baseline.

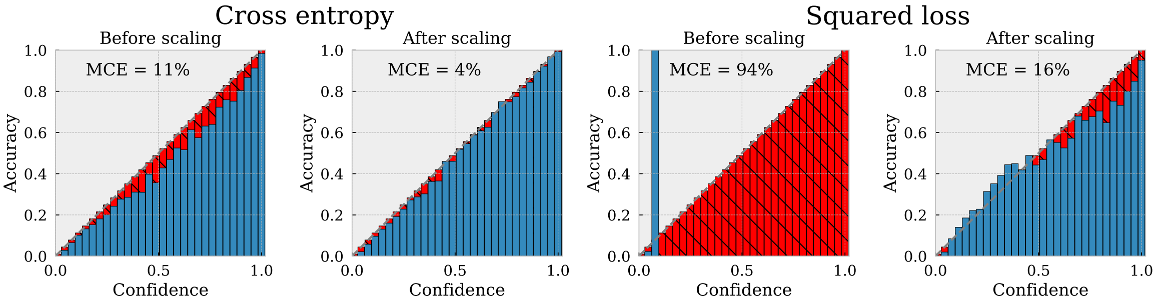


Figure 17. **TCT trained with squared loss is poorly calibrated on CIFAR-100.** Comparing Maximum Calibration Error (MCE) of two loss functions TCT with IID partitions on CIFAR-100 dataset. Red bins show ideal calibration while blue bins show observed calibration.

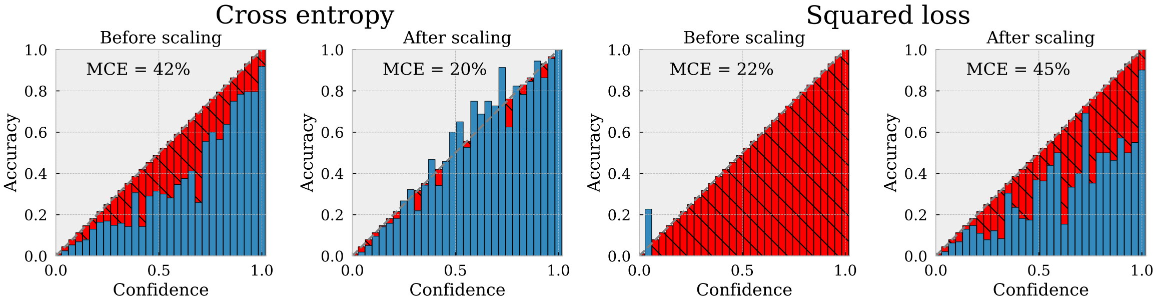


Figure 18. **TCT trained with squared loss is poorly calibrated on Fitzpatrick17k.** Comparing Maximum Calibration Error (MCE) of two loss functions TCT with IID partitions on Fitzpatrick17k dataset. Red bins show ideal calibration while blue bins show observed calibration.